DOCUMENT RESUME

ED 048 908 LI 002 717

AUTHOR Williams, Martha E.
TITLE Provision of Information to the Research Staff.
INSTITUTION Illinois Inst. of Tech., Chicago. Research Inst.
PUB DATE 70
NOTE 19p.; Paper presented at The American Institute of Chemical Engineers 63rd Annual Meeting, Chicago Sheraton Hotel, Nov 29 - Dec 3, 1970

EDRS PRICE EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS Computer Programs, *Data Bases, *Information Dissemination, Information Needs, *Information Retrieval, Information Science, *Information Systems, Information Utilization, *Search Strategies

ABSTRACT

The Information Sciences section at Illinois Institute of Technology Research Institute (IITRI) is now operating a Computer Search Center (CSC) for handling numerous machine-readable data bases. The computer programs are generalized in the sense that they will handle any incoming data base. This is accomplished by means of a preprocessor system which checks for errors and omissions on the supplier tape, converts a data base from the supplier's format to the standard IITRI format. In this way, all tapes submitted are in the IITRI format and the search program can operate on them equally regardless of variation in content that exists from supplier to supplier. IITRI is meeting user needs by providing current awareness and retrospective search services from both document-type and data-type machine-readable files. (MF)

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY

ED048908

PROVISION OF INFORMATION TO THE RESEARCH STAFF

by

Martha E. Williams
Manager
Information Sciences
IIT Research Institute

Paper No. 46C

The American Institute of
Chemical Engineers
63rd Annual Meeting
Chicago Sheraton Hotel

Nov. 29 --- Dec. 3, 1970

002 717

ERIC

Information is both the product of research and an essential ingredient of research. The relationship between researchers and the information they use and generate is a cyclic regenerative relationship.

Provision of required scientific and technical information to researchers has been and will continue to be the responsibility of management. There are, however, many different ways in which this can be done and naturally the needs of researchers in different institutions vary widely depending on whether they conduct basic or applied research and development.

Scientists and engineers are familiar with and use the traditional sources of information. The retrieval and use of information, though, is becoming more problematic because of the monumental volume of information that is being produced in the form of books, journals, documents, etc. Currently more than 2 million scientific and technical papers are published each year and, since the turn of the century, the amount of scientific and technical literature has been doubling every 10-15 years.

Numerous solutions to the information explosion problem have been posed, such as reducing the number of articles published, publishing only summaries or abstracts of articles, or, retaining full documentation on magnetic tape only, and announcing the existence of the information to persons in the appropriate subject areas. I would hate to hazard a guess as to whether such solutions are really forthcoming, but in our "publish or perish"

society where publications effect both salary and ego boosting, I doubt that printed publications will disappear in the near future.

Traditionally the way a scientist has been able to sort out and access the information that is relevant to his interests has been by means of searching the abstracting journals, such as Chemical Abstracts (CA), Biological Abstracts (BA) and Engineering Index. There are some 200 technical societies, governmental agencies and commercial organizations that produce abstract journals or bulletins and searching these publications is no longer a simple job. CA is now publishing approximately 280,000 references per year, BA approximately 250,000 per year and Engineering Index approximately half that amount.

In the past few years some technological changes have taken place that simplify the searching of these large data collections. Many abstracting journals, primary journals and indexes are now being printed by means of computerized typesetting for economy and speed of reproduction. A by-product of this is the existence of machine-readable journals or indexes on magnetic tape and each tape then becomes a searchable data base.

Heretofore one of the largest hindrances to the use of computers for information retrieval has been the very high cost of keyboarding the information to get it into machine-readable form. Fortunately, as more and more journals adopt computerized photocomposition, more machine-readable files will exist. There

are approximately 50 available data bases and most of the major abstracting and indexing organizations produce and lease or sell their data bases to others.

Among the organizations that use computers to aid them in publication of their announcement and abstracting journals or indexes are:

Defense Documentation Center

NASA

AEC

National Technical Information Service (NTIS) (formally Clearinghouse for Federal Scientific and Technical Information)

National Library of Medicine

National Agricultural Library

Chemical Abstracts

Biological Abstracts

American Institute of Physics

American Petroleum Institute

Institute for Scientific Information

American Society for Metals

Engineering Index

Searchable data bases are being produced by all of these organizations and many others.

It is the job of responsible management to keep up to date with new technology that can influence or affect research in their organizations. Most managers of chemical research are

cognizant of modern analytical techniques and new instrumentation, and they may use computers for simulating and monitoring experiments, or to assist them in mathematical modeling. But, how many managers of research are aware of the modern techniques for handling chemical information? There are new information sources and new methods for storing and retrieving the information. I am not implying that every company needs a computerized information storage and retrieval system of its own, whether it be for handling internally generated information or information available through outside sources. In either case, the design, operation and maintenance of a computerized retrieval system is expensive in terms of machine time, personnel, and creation or lease of data bases.

Management should be aware of the new information resources and of the means for accessing them. One way in which these resources have become available to a wide variety of users is through the establishment of computerized information dissemination centers. The Information Sciences section at IIT Research Institute, with funding from the National Science Foundation, has designed and is now operating a Computer Search Center (CSC) for handling numerous machine-readable data bases and meeting user needs by providing the desired sources and services with minimal restrictions and a high degree of flexibility. The users of the Computer Search Center are scientists and engineers in industry, academic institutions and other research organizations.

The center was designed to provide current awareness and retrospective search services from both document-type and data-type machine-readable files. Since there were no available computer search programs that met all of our objectives at the time we began design of our system and since we wanted our software to handle a wide variety of data bases, we developed a new set of generalized, modular, transferable programs that provide considerable flexibility to the user in terms of profile options, search capability, searchable data elements, and output format.

Generalized Programs

The programs are generalized in the sense that they will handle virtually any incoming data base. This is accomplished by means of a preprocessor system. We have written preprocessor programs for each different data bases to be searched. The preprocessor program checks for errors and omissions on the supplier tape, converts a data base from the supplier's format (whether it be CA, BA, Engineering Index, or any other) to the standard IITRI format. In this way all tapes that are submitted to the search program are in the same IITRI format and the search program can operate on them equally regardless of variation in content that exists from supplier to supplier.

Different data base suppliers include different data elements on their tapes. We accomodate this variation by allowing an open-ended number of data elements in our data tape

format. Data elements are specified and identified by type and field length. A code is assigned to each type of data element and some of these are the following:

DATA TYPE	DATA ELEMENT
01	SOURCE INFORMATION (CODEN, JOURNAL REFERENCE, PAGINATION AND DATES)
02	TITLE OF ARTICLE
03	AUTHOR(S)
04	SHORT JOURNAL TITLE
05	KEYWORD(S), INDEX TERMS, or CA SECTION NUMBER
06	REGISTRY NUMBER
07	MOLECULAR FORMULA
08	CORPORATE AUTHOR
09	ABSTRACT OR TEXT
10	BA CROSS CODE
11	BA BIOSYSTEMATIC CODE

Since the number of data elements is open-ended, we can accommodate any data element that exists on any current or future data base.

Modular Programs

The programs are modular in the sense that our software system is made up of a set of modules or discrete blocks for handling separate operations within the system. There are five basic functions provided by the programs: (1) tape format

conversion, (2) profile preparation, (3) search, (4) output generation, and (5) maintenance of statistics. The program for any one of these functions is made up of one or more separate subroutines, and an individual subroutine can be called for, omitted, or changed without disrupting the rest of the system.

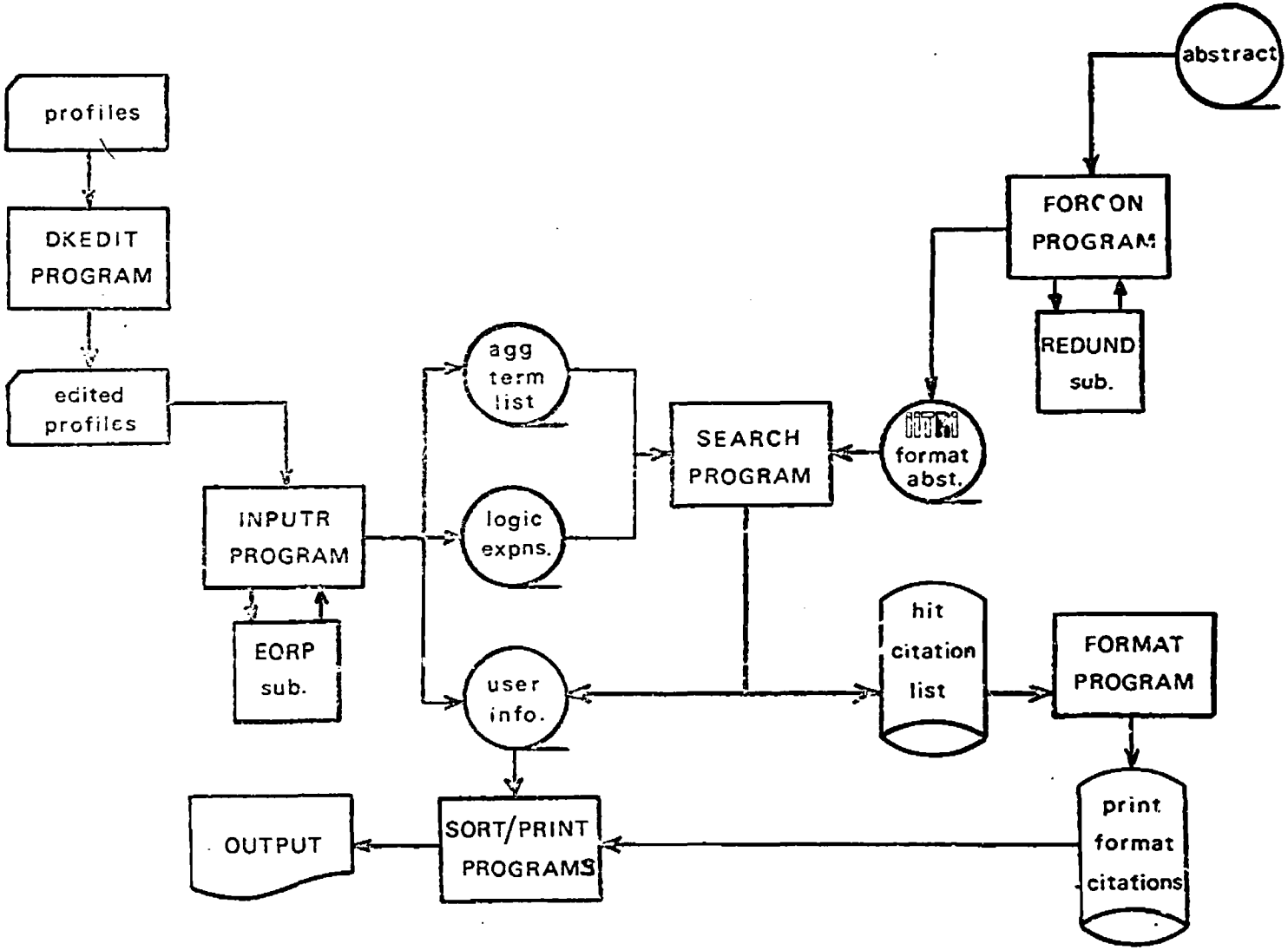
A block diagram of the programming system is shown in Figure 1.

Program Transferability

One of CSC's objectives was the development of programs that could run at a variety of installations. We did not want to be tied to one particular computer configuration for our own use and we also wanted to be able to lease and install the software for other organizations whose volume of searching might warrant operating their own system. Inasmuch as the IBM 360 family of computers represents a large segment of the computer field, we decided to program for the 360. Programs were written to be run on 360's from a model 40 on up. They require a minimum of two tape drives, one or more disks and, assuming approximately 3,000 search terms (200 profiles of 15 terms each), 256K bytes of core storage.

In an effort to achieve machine independence, installation independence and program transferability, the high level compiler language PL/1 was adopted in preference to the more economical Basic Assembly Language. The fact that these

FIGURE 1 - Programming System Block Diagram



objectives have been met is demonstrated by the fact that CSC programs, in both source and object code, were run at seven different computer installations on 360 models 40, 50, 65, 67 and 75, using 2311 and 2314 disc drives with PCP, MFT and MVT processors, under three versions of OS, 15-16, 17 and 18. In no case was any significant problem encountered. Two releases of PL/1 -- 4.3 and 5 -- have been used. PL/1 has proved quite satisfactory because of the facilities it provides for manipulating bit and character strings, handling multi-dimensional arrays and structures, and performing INPUT/OUTPUT operations. IITRI programmers found PL/1 easy to learn and the time required for program development was considerably less than it would have been if Assembly Language had been used.

Flexibility

A search question or interest profile can represent the interests of an individual scientist or a group of scientists. Considerable flexibility is allowed to the user in preparing his profile for computer handling.

Terms

A profile is simply a list of terms representing the user's interest, connected together by logical symbols that indicate the logical relationships between the terms. The terms themselves may be subject terms (text type terms) or any other kind of term that is included on the tape to be searched. For example a search term may be an author name, company name,

journal name (as represented by an ASTM CODEN), CAS registry number representing a unique chemical compound, CA section number, molecular formula, BA CROSS code, BA Biosystematic code, or any other data element. The terms then can be included in the profile either as positive search terms or they may be included as negative or "NOT" terms.

Truncation

Terms are designated with the appropriate truncation mode.

Since many data bases include titles, which are author generated and therefore uncontrolled, it is necessary to include in one's profile all forms of a desired term to ensure retrieval of the desired information. In order to simplify this task of specifying all possible relevant word forms and fragments, CSC has allowed all options in truncation. That is left, right, both and none modes of truncation are permitted. The singular and plural forms of a term can be retrieved by using right truncation e.g., chloride* would retrieve both chloride and chlorides. An example using the term AZO is given in Figure 2.

Figure 2 - Modes of Term Truncation

Mode		Input Format	Action	Examples
None	0	AZO	Retrieves only the term AZO	AZO
Left	1	*AZO	Retrieves any term ending in AZO	AZO, DIAZO, HYDRAZO
Right	2	AZO*	Retrieves any term beginning with AZO	AZO, AZOXY, AZOLE
Both	3	*AZO*	Retrieves any terms in which AZO appears	AZO, DIAZO, HYDRAZO, AZOXY, DIAZOMETHANE AZOLE

Weights

If a user wishes, he may indicate the relative importance of different terms in his profile by assigning different weights to them. For example, if one were interested in information about halogens but were far more interested in bromine than the others, he might assign a higher weight to bromine. Weights can be used to rank output, or to further refine profiles, in which case a threshold weight is assigned so that no citation with a weight below the threshold weight -- even though it may satisfy term matches and logic requirements -- will be a hit item.

Links

Terms that are semantically associated can be linked together in a single expression. That is, several terms that are synonymous terms, related terms, or hierarchically-related broader and narrower terms, can be represented by a single character code (letters of the alphabet from A to Z). This simplifies the user's task of writing his logic expression. He can merely specify a link designator rather than indicate the multiple terms joined by the link in cases where any one of the terms would be equally satisfactory in the logic expression. For example, if a user were interested in halogens he could use the link designator "A" to represent the terms: halogen-halide - fluorine - chlorine - bromine - iodine. In writing his logic expression he would specify the letter "A" in place of writing out the six terms. In effect "A" means halogen or halide or fluorine, or chlorine or bromine or iodine.

Logic

The CSC system is extremely flexible with respect to logic. It uses full free-form Boolean logic. That is, terms can be related to one another by means of the standard Boolean operators AND, OR, and NOT. AND specifies conjunction between two terms or operands. OR specifies disjunction between pairs of terms and NOT is used to preclude the appearance of a term. The symbols used for the Boolean operators are & for AND, | for OR, and ¬ for NOT. The terms or operands can be

connected by means of the operators in any manner that can be represented by a legitimate mathematical parenthetical expression. The CSC system is not restricted to the use of OR within parameters and AND between parameters as many other systems are. The logic system permits nesting to whatever degree the user requires i.e., it can handle parentheses within parentheses within parentheses ....

Profile Example

A hypothetical user's question concerns air pollution legislation. He would like information relating to laws, codes, bills etc. that relate to air pollution or control of air pollution. References to the National Air Pollution and Control Administration (NAPCA) are of particular interest. Articles published by A. Liebman and A. J. Ganner should be omitted and since the hypothetical user regularly reads Environmental Science and Technology (ESTHA), he does not wish to receive any references published in that journal. Relevant references are likely to be found in section 59 of Chemical Abstracts.

Terms used to express this search question are give in Figure 3. All terms are numbered sequentially and associated with each term is its truncation mode (TR), Term Type code number, link code and assigned weight.

In the logic expression for this question terms are represented by their link codes, and they are related to one another by the appropriate logic symbols in mathematical

TERM NUMBER	TR	TERM TYPE	LINK	WT	TERM (Maximum of 20 characters exclusive of asterisk. Use upper case letters - Do not keypunch asterisk)
0.0.1	2	0.2	A	8	LEGISLAT*
0.0.2	2	0.2	A	6	LAW*
0.0.3	2	0.2	A	5	BILL*
0.0.4	2	0.2	A	8	ORDINANCE*
0.0.5	2	0.2	A	5	CODE*
0.0.6	2	0.2	A	5	GUIDELINE*
0.0.7	2	0.2	B	2	ATMOSPHER*
0.0.8	0	0.2	B	5	AIR
0.0.9	0	0.2	B	2	AIRBORNE
0.1.0	2	0.2	C	8	POLLUT*
0.1.1	2	0.2	C	5	CONTAMINA*
0.1.2	2	0.2	C	2	WASTE*
0.1.3	2	0.2	C	2	PARTIC*
0.1.4	2	0.2	C	2	POISON*
0.1.5	2	0.2	C	3	STAGNAT*
0.1.6	2	0.2	C	2	EFFLUEN*
0.1.7	2	0.2	C	0	CORRO*
0.1.8	2	0.2	C	5	PURIF*
0.1.9	2	0.2	C	2	CLEAN*
0.2.0	2	0.2	C	5	IMPURIT*

TERM NUMBER	TR	TERM TYPE	LINK	WT	TERM (Maximum of 20 characters exclusive of asterisk. Use upper case letters - Do not keypunch asterisk)
0.2.1	2	0.2	C	2	DIRT*
0.2.2	0	0.2	D	7	SMOKE
0.2.3	0	0.2	D	7	HAZE
0.2.4	0	0.2	D	5	SMOG
0.2.5	0	0.2	D	7	EXHAUST GAS
0.2.6	0	0.2	D	7	FLUE DUST
0.2.7	0	0.2	D	7	FLUE GAS
0.2.8	0	0.2	D	5	FLY ASH
0.2.9	2	0.2	D	5	FUME*
0.3.0	0	0.2	D	5	SOOT
0.3.1	2	0.2	D	3	VAPOR*
0.3.2	0	0.2	D	9	CA059000
0.3.3	2	0.8	D	9	NAPCA
0.3.4	2	0.3	E	0	LIEBMAN, A*
0.3.5	0	0.3	E	0	GANNER, AJ
0.3.6	0	0.1	E	0	ESTHA

TERM TYPE

CODEN	0 1	CROSS CODE	1 0
TEXT	0 2	BIOSYSTEMATIC INDEX	1 1
AUTHOR	0 3		
REGISTRY NUMBER	0 6		
MOLECULAR FORMULA	0 7		
CORP. AUTHOR	0 8		

NO LINK: LEAVE BLANK
LINK: A-Z

WEIGHT: 0 - 9

SEARCH CODES

TRUNCATION MODE	(TR)
NONE	0
LEFT	1
RIGHT	2
BOTH	3

form.

$$\neg E\&(A\&((B\&C)|D))$$

Output Sort and Print Options

With respect to output, several options are open to the user. For example, he can specify an output limit. This is an upper limit to the number of hit citations he would like to have printed out for him. The user can indicate whether he wants his output printed on cards or paper, and he can specify the way in which the output should be sorted. There are three sort options, alphabetically by first author's last name, numerically in ascending order by reference or abstract number, or in descending weight order. The standard output that is sent to the user is prepared on 5" x 8" cards; however, output can be printed on paper or multilith masters if desired.

Services

The Computer Search Center was designed to provide current awareness and retrospective search services. The CSC has been operational since September 1969 and the principal service we are currently offering is SDI searches against the CAS Condensates and BIOSIS's Biological Abstracts and BIORESEARCH Index on a production basis. We have written format conversion programs for several other data bases and we will provide service from any data base for which we have a sufficient number of subscribers. We are currently conducting a market survey to determine the degree of interest there is in various data bases.

Retrospective searches are available on both a manual and machine basis. To date very few data bases date back very far. CAS Condensates has been in existence only two years and many other data bases are even newer than that. Until retrospective data files exist the majority of retrospective searches will continue to be done by hand. The tapes that have been used for the SDI searches are arranged in serial form and so are not amenable to economical retrospective search. We are currently conducting a research program to investigate the feasibility of merging multiple data bases and to investigate methods of inverting and compressing large data bases to permit economical retrospective searches.

In addition to the services described above we have developed software for the creation, and maintenance of personal disk or tape libraries. Traditionally the researcher scans or searches hard copy documents: (1) technical publications such as abstracting journals, government reports and primary journals, (2) internal reports such as technical reports and laboratory reports or (3) product information such as vendor data or company formulations. When he finds something of interest that he may want to recall sometime in the future for writing a report or technical article, he either notes it mentally or records the bibliographic information on a card or piece of paper for his personal file. He then files the reference in some logical order or merely stashes it away and hopes he'll be able to find it when he needs it.

An alternative to this method is to maintain these references on a personal disk or tape file where they can be inexpensively stored and retrieved at the desired time. We have developed software for creating personal computerized files by collecting a user's SDI output and storing it for him. The user can then delete references that are of no interest and maintain those he has judged relevant. If the user is running his profile against several data bases the output from all of them can be stored for him in the same format. The user can also add to his file additional index terms and codes or any references from other sources such as company reports, vendor literature or journals that are not covered by the SDI data bases. Updating from the SDI system is done automaticaly and additions to the file from other sources can be done at the user's convenience. Searching is done when the user wants it and output can be provided on cards or in bibliographic form. A personal disk library contains only the information that is pertinent to the users interests and his judgments have been made about all items in the file. It is a pre-screened tailor-made data base.

Conclusion

There are many new data bases and sources of information

relevant to the interest of researches. There are also many new services and organizations whose raison d'etre is to provide information from the new data bases to users. If management is to continue to maintain responsibility for providing needed information to staff members in the most economic and efficient manner, it behooves management to become familiar with the new sources and services in order to make informed judgements.